Supplementary Material for "Distributional Information Embedding: A Framework for Multi-bit Watermarking"

A. *Proof of Lemma 1*

Proof. Let $P_e = \Pr(\hat{M} \neq M)$. From the Fano's inequality, we have

$$\mathsf{H}(M|\hat{M}, \zeta_1^T) \leq \mathsf{H}(M|\hat{M}) \leq 1 + P_e \log m.$$

The entropy of M is upper bounded by

$$\begin{aligned}\log m = \mathsf{H}(M) = \mathsf{H}(M|\zeta_1^T) &= \mathsf{I}(M; \hat{M}|\zeta_1^T) + \mathsf{H}(M|\hat{M}, \zeta_1^T)\\ &\leq \mathsf{I}(M; X_1^T|\zeta_1^T) + 1 + P_e \log m\\ &\leq H(X_1^T|\zeta_1^T) + 1 + P_e \log m,\end{aligned}$$

which leads to

$$\frac{\log m}{T} \leq \frac{H(X_1^T|\zeta_1^T)}{T} + \frac{1}{T} + P_e \frac{\log m}{T}.$$

If $P_e \to 0$ as $T \to \infty$, we have

$$\frac{\log m}{T} \leq \frac{H(X_1^T|\zeta_1^T)}{T} \leq \mathsf{H}(P_X) \leq \sup_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X).$$

□

B. *Proof of Lemma 2*

Proof. For any $i \neq j$, define the relative entropy typical set

$$\mathcal{A}_{\epsilon,i,j}^{(T)}(\mathbb{P}_i \| \mathbb{P}_j) := \left\{ (x_1^T, \zeta_1^T) : \left| \frac{1}{T} \log \frac{\mathbb{P}_i(x_1^T, \zeta_1^T)}{\mathbb{P}_j(x_1^T, \zeta_1^T)} - \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) \right| \leq \epsilon \right\}.$$

We have $\mathbb{P}_j(\mathcal{B}_{T,j}^{\mathrm{c}}) = 1 - \mathbb{P}_j(\mathcal{B}_{T,j})$ and

$$\begin{aligned}\mathbb{P}_j(\mathcal{B}_{T,j}) &= 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i}) \leq 1 - \sum_{i:i\neq j} \mathbb{P}_j(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)})\\ &\leq 1 - \sum_{i:i\neq j} \sum_{(x_1^T, \zeta_1^T) \in \mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}} \mathbb{P}_i(x_1^T, \zeta_1^T) \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\\ &= 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)})\\ &\overset{\text{(a)}}{\leq} 1 - \sum_{i:i\neq j} \exp(-T(\mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))(1 - 2\epsilon)\\ &\leq 1 - m(1 - 2\epsilon) \exp(-T(\min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\\ &\leq 1 - m(1 - 2\epsilon) \exp(-T(\max_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \min_{i:i\neq j} \mathsf{D}_{\mathsf{KL}}(P_{X,\zeta|M=i} \| P_{X,\zeta|M=j}) + \epsilon))\end{aligned}$$

where (a) follows since $\mathbb{P}_i(\mathcal{B}_{T,i} \cap \mathcal{A}_{\epsilon,i,j}^{(T)}) = 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}} \cup (\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - \mathbb{P}_i(\mathcal{B}_{T,i}^{\mathrm{c}}) - \mathbb{P}_i((\mathcal{A}_{\epsilon,i,j}^{(T)})^{\mathrm{c}}) \geq 1 - 2\epsilon$ for sufficiently large T. The proof is thus complete. □

C. *Proof of Theorem 3*

Let $\eta = T^{-\frac{1}{4}}$ and define the set $\mathcal{A}_{\eta,j}^{(T)}$ of jointly typical sequences $\{(x_1^T, \zeta_1^T)\}$ w.r.t. the distribution $P_{X,\zeta|M=j}$ as

$$\begin{aligned}\mathcal{A}_{\eta,j}^{(T)} := \Bigg\{ (x_1^T, \zeta_1^T) \in \mathcal{X}^T \times \mathcal{Z}^T : &\left| -\frac{1}{T} \log P_X^T(x_1^T) - \mathsf{H}(P_X) \right| \leq \eta, \left| -\frac{1}{T} \log P_\zeta^T(\zeta_1^T) - \mathsf{H}(P_\zeta) \right| \leq \eta,\\ &\left| -\frac{1}{T} \log P_{X,\zeta|M=j}^T(x_1^T, \zeta_1^T) - \mathsf{H}(P_{X,\zeta|M=j}) \right| \leq \eta \Bigg\}.\end{aligned}$$

First, we bound the probability of the atypical sets $(\mathcal{A}_{\eta,X}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,\eta}^{(T)})^{\mathrm{c}}, (\mathcal{A}_{\eta,j}^{(T)})^{\mathrm{c}}$. From the union bound, we have

$$\begin{aligned}\mathbb{P}_j((X_1^T, \zeta_1^T) \notin \mathcal{A}_{\eta,j}^{(T)}) \leq \mathbb{P}_j\left(\left| -\frac{1}{T} \log P_X^T(x_1^T) - \mathsf{H}(P_X) \right| \geq \eta \right) + \mathbb{P}_j\left(\left| -\frac{1}{T} \log P_\zeta^T(\zeta_1^T) - \mathsf{H}(P_\zeta) \right| \geq \eta \right)\\ + \mathbb{P}_j\left(\left| -\frac{1}{T} \log P_{X,\zeta|M=j}^T(x_1^T, \zeta_1^T) - \mathsf{H}(P_{X,\zeta|M=j}) \right| \geq \eta \right).\end{aligned}$$

Then, by the Chernoff bound, we have

$$\begin{aligned}\mathbb{P}_j\bigg(\bigg|-\frac{1}{T}\log P_X^T(x_1^T)-\mathsf{H}(P_X)\bigg|\geq\eta\bigg) &\leq 2\mathbb{P}_j\bigg(-\frac{1}{T}\log P_X^T(x_1^T)-\mathsf{H}(P_X)\geq\eta\bigg)\\ &\leq 2\exp\bigg(-T\sup_{s\geq 0}(s\eta-\log\mathbb{E}[\exp(-s\log P_{X_1^T}(X_1^T))])\bigg)\\ &\overset{\text{(a)}}{\approx} 2\exp\bigg(-T\sup_{s\geq 0}(s\eta-\big(-s\mathbb{E}[\log P_{X_1^T}(X_1^T)]+s^2\mathbb{E}[(\log P_{X_1^T}(X_1^T))^2]\big)\bigg)\\ &\overset{(b)}{=} 2\exp(-\Omega(T\eta^2))=\exp(-\Omega(T^{\frac{1}{2}})),\end{aligned}$$

where (a) follows from the Taylor expansion of $\exp(\cdot)$ and $\log(\cdot)$ and (b) follows since the maximum is achieved by $s=O(\eta)$. The rest of the terms in the union bound can be similarly proved.

Thus, the probability of the atypical set is upper bounded by

$$\mathbb{P}_j((X_1^T,\zeta_1^T)\notin\mathcal{A}_{\eta,j}^{(T)})\leq 3\exp(-\Omega(T^{\frac{1}{2}}))=\exp(-\Omega(T^{\frac{1}{2}})).$$

Let $P_X^*=Q_X$, $\mathcal{Z}\subset\mathbb{Z}$ and design $P_\zeta^*\in\mathcal{P}(\mathcal{Z})$ such that $\mathsf{H}(P_\zeta^*)=\mathsf{H}(P_X^*)$.

For any $\gamma^*\in\Gamma^*$, any $j\in[m]$, the j-th error probability is given by

$$\begin{aligned}\beta_j(\gamma^*,P^*_{X_1^T,\zeta_1^T|M=j}) &= \sum_{x_1^T,\zeta_1^T}P^*_{X_1^T,\zeta_1^T|M}(x_1^T,\zeta_1^T|j)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)\neq j\}\\ &\leq \sum_{(x_1^T,\zeta_1^T)\in\mathcal{A}_{\eta,j}^{(T)}}P^*_{X_1^T,\zeta_1^T|M}(x_1^T,\zeta_1^T|j)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)\neq j\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= \exp(-\Omega(T^{\frac{1}{2}}))\to 0 \text{ as } T\to\infty.\end{aligned}$$

For $j=0$, the worst-case false alarm error probability is upper bounded as follows. For any $x_1^T\in\mathcal{X}^T$,

$$\begin{aligned}\sum_{\zeta_1^T}P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)\neq 0\} &\leq \sum_{\zeta_1^T\in\mathcal{A}_{n,\zeta}^{(T)}}P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)\neq 0\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &\doteq \sum_{i\in[m]}\sum_{\zeta_1^T\in\mathcal{A}_{n,\zeta}^{(T)}}e^{-T\mathsf{H}(\zeta)}\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)=i\}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= me^{-T\mathsf{H}(\zeta)}+\exp(-\Omega(T^{\frac{1}{2}}))\\ &= \alpha+\exp(-\Omega(T^{\frac{1}{2}}))\\ &\xrightarrow{T\to\infty}\alpha.\end{aligned}$$

Since any distribution Q_X^T can be written as a linear combinations of $\delta_{x_1^T}$, we have

$$\sup_{Q_X}\beta_0(\gamma^*,Q_X\otimes P_\zeta^*)=\sup_{Q_X}\sum_{x_1^T,\zeta_1^T}Q_X^T(x_1^T)P_\zeta^*(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)\neq 0\}\leq\alpha$$

D. *Proof of Theorem 4*

First, we have

$$\beta_j(\gamma,P_{X_1^T,\zeta_1^T|M=j})=\sum_{i:i\neq j}\mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i).$$

For any $i\neq j$, the optimization constraints imply that for any $y_1^T\in\mathcal{X}^T$,

$$\alpha\geq\sup_{P_{X_1^T,\zeta_1^T|M=i}}\beta_i(\gamma,P_{X_1^T,\zeta_1^T|M=i})\geq\sum_{\zeta_1^T}P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{\gamma(y_1^T,\zeta_1^T)\neq i\}.$$

Then we have

$$\begin{aligned}\mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)\neq i) &= \sum_{x_1^T,\zeta_1^T}P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\}\\ &\overset{(a)}{\leq}\sum_{x_1^T}(P_{X_1^T}(x_1^T)\wedge\alpha),\end{aligned}$$

where (a) follows since $\sum_{\zeta_1^T}P_{\zeta_1^T}(\zeta_1^T)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\}\leq\alpha$ and $\sum_{\zeta_1^T}P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq i\}\leq\sum_{\zeta_1^T}P_{\zeta_1^T}(\zeta_1^T)P_{X_1^T|\zeta_1^T,M=j}(x_1^T|\zeta_1^T,M=j)=P_{X_1^T}(x_1^T)$ for all x_1^T.

Consequently,

$$\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) = \sum_{i:i\neq j} \mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i) \geq \sum_{i:i\neq j}(1-\sum_{x_1^T}(P_{X_1^T}(x_1^T)\wedge\alpha)) = m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+$$
$$\geq \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\leq d} m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+,$$

where m,α should satisfy $m\sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+ \leq 1$ and the lower bound holds for all γ and $P_{X_1^T,\zeta_1^T|M}$.

Additionally, the analyses still hold when $P_{\zeta_1^T|M=j}$ are not the same for all j.

E. *Proof of Theorem 5*

Choose $\mathcal{Z}\subset\mathbb{Z}^T$ such that $|\mathcal{Z}|^T = m|\mathcal{X}|^T+1$.Randomly pick a sequence $\tilde{\zeta}_1^T\in\mathcal{Z}^T$ and partition $\mathcal{Z}^T\backslash\{\tilde{\zeta}_1^T\}$ into m disjoint subsets $\{\mathcal{S}_j\}_{j=1}^m$ of equal size. Define a set of decoders as

$$\Gamma_{\tilde{\zeta}_1^T} := \left\{\gamma \;\middle|\; \begin{array}{c} \gamma(x_1^T,\zeta_1^T) = \begin{cases} j, & \text{if } \zeta_1^T \neq \tilde{\zeta}_1^T \\ & \text{and } x_1^T = h_j(\zeta_1^T), \, , \\ 0, & \text{otherwise}, \end{cases} \\ \text{for some group of bijective functions } \{h_j:\mathcal{S}_j\to\mathcal{X}^T\}_{j=1}^m \end{array}\right\}$$

For any $\gamma\in\Gamma_{\tilde{\zeta}_1^T}$, under the watermarking scheme presented in Theorem 5, we have:

– For any $j\in[m]$, the j-th error probability is give by

$$\beta_j(\gamma, P_{X_1^T,\zeta_1^T|M=j}) = \sum_{i\in[0:m]\backslash j} \mathbb{P}_j(\gamma(X_1^T,\zeta_1^T)=i)$$
$$= m \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\leq d} \sum_{x_1^T}(P_{X_1^T}(x_1^T)-\alpha)_+.$$

– False alarm error: for any $x_1^T\in\mathcal{X}^T$,

$$\sum_{\zeta_1^T} P_\zeta(\zeta_1^T)\mathbb{1}\{\gamma(x_1^T,\zeta_1^T)\neq 0\} = \sum_{i=1}^m\sum_{\zeta_1^T} P_\zeta(\zeta_1^T)\mathbb{1}\{\gamma^*(x_1^T,\zeta_1^T)=i\}$$
$$= (P^*_{X_1^T}(x_1^T) - m(P^*_{X_1^T}(x_1^T)-\alpha)_+) + (m-1)(P^*_{X_1^T}(x_1^T)-\alpha)_+$$
$$= P^*_{X_1^T}(x_1^T) - (P^*_{X_1^T}(x_1^T)-\alpha)_+$$
$$= P^*_{X_1^T}(x_1^T)\wedge\alpha \leq \alpha.$$

Since any distribution $Q_{X_1^T}$ can be represented by a linear combination of $\delta_{x_1^T}$, the worst-case false alarm error is upper bounded by

$$\sup_{Q_{X_1^T}} \beta_0(\gamma, P_{X_1^T,\zeta_1^T|M=j}) \leq \alpha.$$